述评

大数据:微生物组学及其他生物医学领域的机遇与挑战

徐振江
科罗拉多大学生物前沿研究所,美国 博尔德 80303

摘要:随着高通量技术的发展,生物数据大爆发式地增长。如何有效地利用生物大数据成为现代生物学的机遇和挑战。大数据和传统数据相比,呈现出很多不同的特点,包括常被提到的3个v (volume, variety, velocity 即数据量的巨大、数据类型的多样和数据采集和处理的快速)。本文针对生物医学研究,详细介绍了大数据的杂乱性、可重复利用性、开放性等几个特点。同时结合微生物组学在元分析方面的最新进展,并用实例来阐述了我们在大数据采集方面应该有前瞻性的考虑,提出了在数据管理上如何保护隐私的挑战,探讨了对大数据进行分析的工具和方法。
关键词:大数据;微生物组学;生物医学;高通量技术;多样性;杂乱性;可重复利用性;开放性

Big Data: the great opportunities and challenges to microbiome and other biomedical research

XU Zhenjiang
Biofrontiers Institute, University of Colorado, Boulder, CO 80303, USA

Abstract: With the development of high-throughput technologies, biomedical data has been increasing exponentially in an explosive manner. This brings enormous opportunities and challenges to biomedical researchers on how to effectively utilize big data. Big data is different from traditional data in many ways, described as "3Vs" - volume, variety and velocity. From the perspective of biomedical research, here I introduced the characteristics of big data, such as its messiness, re-usage and openness. Focusing on microbiome research of meta-analysis, the author discussed the prospective principles in data collection, challenges of privacy protection in data management, and the scalable tools in data analysis with examples from real life
Key words: Big Data; microbiome; biaomedical; high-throughput technology; variety; messiness; re-usage; openness

随着信息化时代的到来,我们在生活的方方面面都积累了大量的数据。所谓大数据(Big Data),不仅仅是信息量的巨大,同时也是信息的复杂性和多样性。相对于传统的抽样数据,在信息时代以前,由于采样的困难、计算机技术或者分析手段的限制,我们通常无法收集每个个体的数据,只能在总体(population)里进行抽样(sampling),通过分析这些样本,进而推测总体的特征。在大数据时代,我们往往把所有个体的各方面的信息都进行收集整合,得出意想不到的结论。大数据分析已经广泛应用于许多商业、社会科学和自然科学领域。例如,通过记录顾客以往购买的产品和浏览的网页,甚至是浏览者鼠标指向的产品链接,亚马逊能够推测浏览者可能感兴趣的产品,并推荐给消费者。众所周知,情绪是可以通过人与人的接触(行为、表情、言语)互相传染,但最近研究通过分析Facebook上面海量的数据,发现即使在互联网上,情绪也可以通过社交媒体传播开来[1]。

收稿日期:2015-01-08
作者简介:徐振江,University of Rochester获得博士学位,现于美国科罗拉多大学进行博士后工作。开发RNA的结构预测的新算法,并结合机器学习和ncRNA结构,预测人基因组和微生物基因组中的新ncRNA。结合计算机与微生物组学分析人体和环境微生物群落,发现microbial signature,用于医学临床检测和治疗。近3年来,共发表SCI论文12篇

在生物医学方面,随着高通量技术的开发,例如microarray、新一代质谱和测序平台的出现,我们能够方便地获取大量的各种组学(omics)的数据,使得各大数据库呈指数级的增长。许多国际合作项目,比如千人基因组计划(The 1000 Genomes Project)[2],癌症基因组图谱(The Cancer Genome Atlas, TCGA))[3],人类微生物组计划(Human Microbiome Project, HMP)[4-5],人类肠道宏基因组计划(Metagenomics of the Human Intestinal Tract consortium, MetaHIT)[6]都产生了大量的数据,再也不局限在几个蛋白分子或几段DNA序列之内。那么,这些大数据都有什么特点? 它能给生物医学研究带来什么帮助? 我们应该如何采集、管理、分析生物大数据? 本文将尝试着一一回答这些问题。

1 重复利用性

在大数据时代,一组数据常常可以在不同的场合重复利用。在微生物组学的研究中,有不少关于炎症性肠病(IBD)或者肥胖症和肠道微生物之间相关性的报道,后续的研究者就可以把所有这些报道的数据整合起来做元分析(meta-analysis),从而发现其中的共同特征,弥补单个研究中由于数据不足或者偏差而可能得出

的缺乏普适性的结论[7]。数据,不像其他一般的商品,它的内在价值不会因为多次重复使用而降低。考虑到将来可能的用处,我们在收集数据的时候,应该尽可能地前瞻性地收集更多的数据。虽然有些信息跟当下的研究关系不大,但是由于数据采集的边际成本比较低,你可以随手收集可能对将来非常重要的数据。在一项研究全球微生物分布规律的生态学研究中,正是由于每组数据都采集了pH、温度等各个环境的理化信息,才最终通过元分析发现盐度是影响微生物分布的最重要的因子[8]。而且,随着数据存储成本的急剧下降,不用担心数据过多而无法保存。运用前瞻性的原则采集数据是大数据时代的一个重要原则。比如,每次疾病患者或者健康人群来医院看病或者体检的时候,我们可以记录下所有方便可测的生理、心理数据,建立纵向的时间序列数据库,为将来提高疾病的诊断和预防技术提供大量的数据支持。由于DNA测序成本的降低和自动化机械手臂在实验操作上的使用,我们甚至可以同时低成本地收集大量的口腔、粪便、皮肤等微生物样本,为将来深入研究微生物跟各种慢性疾病的发病机制之间的关系提供数据。

2 杂乱性

当我们获取大量数据的时候,不可避免地会包含一些不精确或者含有误差的数据。当数据量比较小的时候,我们能够,也应该保证每个数据的精确性,因为在数据少的情况下,任何误差会对结果造成明显的影响。数据精确度和信息量之间,在很大程度上其实是一个权衡关系。随着数据量的增长,我们很难去逐个检查验证每个数据,这将会耗费巨大的人力成本。虽然我们可以通过算法自动寻找可能有问题的数据,但效果非常有限。那么,在以牺牲精确度为代价的情况下,大数据是不是还有意义呢?答案是肯定的。其中最直接的一个例子就是谷歌的机器翻译。在20世纪90年代,用来创建机器翻译模型的数据主要来自于双语的官方文件,这种语料库的质量是非常高的,但是机器翻译的质量一般,无论如何改进算法,效果都十分有限。直到2006年,谷歌抓取互联网上所有的双语材料用来建模,虽然这些语料的质量良莠不齐,但是巨大的数据量大大提高了机器翻译的准确度[9]。

另外一个例子,我们实验室正在开展的一个面向公众的大型科学项目——American Gut Project。每位参与者只要捐赠 $ 99,就可以用我们提供的试剂盒采集自己感兴趣的微生物样本,寄回给我们测序、分析,然后我们再把分析结果以简单易懂的方式返还给他们。这个过程中有很多我们没法控制的因素,数据没法做到完全精确。比如,样本在邮寄的过程中有很多微生物不可避免的进行繁殖,从而可能改变原来的群落结构。确实,我们发现有几个变形菌门(proteobacteria)的细菌进行了大量繁殖。再比如,参与者在填写样本相关的信息(通常称之为meta data)的时候,难免出现描述不准确甚至错误的情况。虽然有诸多影响数据准确度的因素存在,但我们还是能得出一些在数据量小的时候没法得出的、有意义的结论(还未发表)。之所以如此,是因为大数据能够给我们描绘一副全面的图景,即使这幅图景存在一些局部误差,只要这些误差是随机的,就不会太影响我们对整体的认识。

大数据的杂乱性的还有另外一层意思。我们常常把不同来源和不同类型的几组数据整合在一起分析。例如,医疗大鳄Permanente曾经通过整合各个医疗机构的数据和临床病人的电子病历,成功降低病人的就诊率,从而减少了大量医疗成本[10]。但在很多情况下,几组相关的数据是不完全相对应的,杂乱的。因此,数据整合需要对数据进行规范,使之易于分享、合并。数据的规范化不仅仅是指数据的电子信息化和格式的统一,更是指描述语言的标准化。Genomic Standards Consortium正是出于此目的而建立的[11]。该组织在微生物组学领域设立了MIMARKS等标准[12-13],来促进微生物组学数据的交换共享。

3 开放性

基于以上原因,数据整合带来的益处远远高于单组数据,数据共享往往能创造共赢的局面,对所有的数据分享者都是有利的。因此,数据的开放性是大势所趋。近年来,许多国家都意识到这点,极力推动数据的共享。比如,美国从2009年开始要求所有联邦机构必须公开它们收集的公众数据,普通大众可以在 http://www.data.gov/ 自由下载并使用包括农业、金融、能源、科研等各个部门在内的、超过138 470组的数据。英国甚至成立了一个新的研究所Open Data Institute来鼓励数据的共享和使用。欧盟和其他国家(包括澳大利亚、巴西、智利等)也都制定了相应法规和策略,希冀在大数据时代占有一席之地。中国,作为世界上人口最多的国家,在人类健康方面,拥有着巨大的数据资源;同时,中国领土面积广大,土地、海洋、生物的多样性都是许多国家无法比拟的,利用好这些数据可以极大地促进科学技术的发展和社会经济的增长。

4 隐私保护

隐私权无疑是每个公民的基本权利。在大数据时代,如何保护数据来源人的隐私是一个巨大的挑战。隐私保护通常有两驾马车:(1)个人能够选择同意或不同意提供数据;(2)数据的匿名化和去标识化。但针对于

大数据,这两种手段都难以发挥作用。前文提到,数据共享是有效利用大数据的必经之路。虽然数据的原始采集机构可以经过个人的授权同意,但是当要把数据分享给第三方时,就难以再回过去一一争取每个人的授权了。而且,在数据采集的时候,谁是将来的第三方往往是未知的,因此,无法提前将第三方加入到授权条款当中。比如,医院在经过病人同意采集了病人样本,多年以后医院想把所有这些样本共享给另外一家科研机构或者健康保险机构,但是由于样本量的巨大,重新联系原来的每个病人是一件成本巨大、甚至是不可能完成的事情。尽管如此,如果我们能够将样本匿名化,也可以同样保护病人隐私 。但是,大数据是如此地具有个人特征,即使完全地去匿名化,你还是可以通过数据来找到数据的主人。比如,Netflix公司曾公布过一批用户的观影记录,巨额悬赏能改进其影片推荐系统的算法。尽管公布的数据中所有用户的信息都已经仔细地去标识化了,但是,根据用户对不同影片的喜好,通过对比IMDB(the Internet Movie Database)数据库,研究人员仍然有很高的概率来识别用户。这个问题在生物医学研究中更加严重,因为每个人的生物性状更加独特、可识别,健康保险公司很容易把你和你的基因组对应起来,通过分析你基因组来计算你的将来的疾病风险,从而不公平地差别对待每个受保人。不仅每个人的基因组是独特的,每个人所携带的微生物群落也是显著不同的[14],加之越来越多的研究证明了微生物群落跟人健康之间的关系[7, 15-16],因此人体相关的微生物组学也同样面临着隐私保护的问题。

5 云存储和云计算

随着数据量的增长,在本地存储和分析数据对计算机硬件的要求越来越大。云存储和云计算对此提供了切实可行的解决方案。现在国内外许多厂商都有云服务,例如阿里巴巴的阿里云,亚马逊的AWS等。用户只需要购买相应的使用时间,不用担心软硬件的配置,就可以在云端服务器进行各种大型运算。另外,云还解决了数据传输的问题。当数据在云端服务器上时,数据的各方使用者就不需要下载至各自的本地服务器上进行分析。虽然有支持断点续传、快速传输技术(例如fasp,传输速度可达700~800 Mbps),但下载数据仍然要占据大量的带宽。因此,在云端存储数据和进行分析是更佳的选择。

6 大数据的分析

大数据给科学研究提供了全新的思路。传统的研究常常假设驱动,即根据已知的科学事实,对所研究的自然现象提出推测和假说,再通过设计实验来验证假说的成立与否。而面对大数据时,数据驱动的研究方式开始越来越普遍,即不提出任何假说,让数据来引导我们得出科学结论。这在有些研究领域里特别重要。比如,如果我们要通过设计实验来研究一个菌种在群落中的作用,我们就应该把该菌种从群落中剔除,再把该群落接种到无菌的研究对象中去,去检测群落的变化,这在实验上是很难做到的。不像遗传分析,在研究一个基因的功能时,我们可以通过敲除该基因,来观察生物性状的改变是否符合提出的假说。在这种情况下,大数据就可以帮助解答这个问题——我们可以记录下微生物组在不同环境下的分布以及它对各种干预所产生的变化,当这些数据积累地越来越多时,我们就能梳理出哪些细菌在整个微生态系统中可能具有什么样子的功能。

如何从纷繁复杂的大数据中得出有用的结论呢?这就需要包括机器学习和数据挖掘在内的一系列多重变量分析方法(multivariate analyses)。常见的方法有分类分析(classification)、回归分析(regression)、聚类分析(clustering)、主成分分析(principal components analysis,PCA)等。这些方法都广泛应用于生物医学研究当中[17-18]。比如,通过采集尸体上的微生物样本,我们可以根据微生物群落的演替建立回归分析模型,来准确地预测尸体的死亡时间,这将对刑侦提供重要的信息[19]。再比如,运用分类分析,我们可以在整个基因表达谱或者分子谱当中筛选癌症的分子标记,从而对癌症类型做出准确的诊断[20],以制定个性化的治疗方案。对这些分析方法感兴趣的读者可以参考文章后面列出来的资源。

由于数据量的巨大,有些过去适用于抽样数据的统计方法和分析工具缺乏可扩展性(scalability),难以满足我们大数据快速分析的需求。而Hadoop[21]和MapReduce[22]采用分布式系统,高效地利用大型计算机群进行并行运算,大大加快了海量信息的处理。需要提醒的是,大数据得出来的结论,和所有的结论一样,不可盲目相信,应本着科学精神,让实验和时间来检验。例如,Google早在2008年就成立了Google Flu Trend服务,通过汇总搜索数据,统计各个地方流感相关的搜索关键词,来预测流感疫情[23]。但2013年发现Google的预测结果与美国疾病预防控制中心的监测报告相比,严重高估了流感发病率[24]。

虽然随着算法的改进,计算机越来越聪明,但是,机器学习和人工智能目前还是无法和人脑比拟。Foldit[25]是一款让玩家去预测和设计蛋白结构的游戏 ,在短短3周里,Foldit玩家就破译M-PMV逆转录病毒蛋白酶的晶体结构。这个酶在艾滋病毒复制和成熟过程中起着关键的作用,它晶体结构的解析曾困扰计算生物学家十年时间,但是缺乏生化知识的在线的玩家们利用人脑的

空间推理能力一起合作,解决了这个计算机难以解决的问题[26]。同样,EtRNA[27]也是试图通过玩家的参与来寻找RNA分子结构设计的规律。可见,尽管大数据如此重要,但也不是万能的,人类的想像能力、推理能力都有着不可忽视的作用。

最后,需要补充的是,传统的严谨的数据收集方法有其自身适用的地方,大数据不可能完全取代,二者相互辅助,在不同的场合发挥不同的作用。很多生物学领域,都有这大数据的用武之地,等待着我们去发现和挖掘。尤其是随着移动互联网的发展,人们可以在智能手机和其他终端(例如智能血压仪、血糖测试仪等)上对自己的身体状况自测自检,进一步降低大数据的成本。这些健康监测数据将帮助我们在未来给每个人提供个性化的靶标治疗。因此,只要我们培养好自己大数据的头脑和嗅觉,学会让数据说话,大数据在生物医学领域的前景将非常广阔。

参考文献:

[1] Kramer ADI, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks[J]. Proc Natl Acad Sci, 2014, 111(24): 8788-90.

[2] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing[J]. Nature, 2010, 467 (7319): 1061-73.

[3] The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project[J]. Nat Genet, 2013, 45(10): 1113-20.

[4] Turnbaugh PJ, Ley RE, Hamady M, et al. The human microbiome project[J]. Nature, 2007,449 (7164): 804-10.

[5] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome[J]. Nature, 2012, 486 (7402): 207-14.

[6] Nelson KE. Metagenomics of the Human Body[M/OL]. Springer New York; 2011 [2013-10-29]. http://link.springer.com/chapter/10.1007/978-1-4419-7089-3_15

[7] Walters WA, Xu Z, Knight R. Meta-analyses of human gut microbes associated with obesity and IBD[J]. FEBS Lett, 2014,(22)588: 4223-33.

[8] Lozupone CA, Knight R. Global patterns in bacterial diversity[J]. Proc Natl Acad Sci, 2007,104(27):11436-40.

[9] Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data[J]. IEEE Intell Syst, 2009:24, 8-12.

[10]Chen C, Garrido T, Chock D, et al. The kaiser permanente electronic health record: Transforming and streamlining modalities of care[J]. Health Aff (Millwood), 2009, 28(2): 323-33.

[11] Field D, Amaral-Zettler L, Cochrane G, et al. The genomic standards consortium[J]. PLoS Biol, 2011, 9, e1001088.

[12] Yilmaz P, Kottmann R, Field D, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications[J]. Nat Biotechnol, 2011,29(5): 415-20.

[13] Glass EM, Dribinsky Y, Yilmaz P, et al. MIxS-BE: a MIxS extension defining a minimum information standard for sequence data from the built environment[J]. ISME J, 2014, 8(1): 1-3.

[14] Fierer N, Lauber CL, Zhou N, et al. Forensic identification using skin bacterial communities[J]. Proc Natl Acad Sci, 2010, 107(14): 6477-81.

[15] Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes[J]. Nature, 2012, 490(7418): 55-60.

[16] Delzenne NM, Cani PD. Interaction between obesity and the gut microbiota: relevance in nutrition[J]. Annu Rev Nutr, 2011, 31: 15-31.

[17] Knights D, Costello EK, Knight R. Supervised classification of human microbiota[J]. FEMS Microbiol Rev, 2011, 35(2): 343-59.

[18] Xu Z, Malmer D, Langille MGI, et al. Which is more important for classifying microbial communities: who's there or what they can do? [J]. ISME J, 2014, 8(12): 2357-9.

[19]Metcalf JL, Wegener Parfrey L, Gonzalez A, et al. A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system[EB/OL]. http://elife.elifesciences.org/content/2/e01104

[20] Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures[J]. Proc Natl Acad Sci, 2001, 98(26): 15149-54.

[21] Shvachko K, Kuang H, Radia S, et al. The hadoop distributed file system[C]//IEEE 26th Symposium on Mass Storage Systems and Technologies, 2010: 1-10.

[22] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters[J]. Commun ACM, 2008, 51: 107-13.

[23] Ginsberg J, Mohebbi MH, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2009, 457(7232): 1012-4.

[24] Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis[J]. Science, 2014, 343 (6176): 1203-5.

[25] Cooper S, Khatib F, Treuille A, et al. Predicting protein structures with a multiplayer online game[J]. Nature, 2010, 466(7307): 756-60.

[26] Khatib F, DiMaio F, Group FC, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players[J]. Nat Struct Mol Biol, 2011, 18(10): 1175-7.

[27] Lee J, Kladwang W, Lee M, et al. RNA design rules from a massive open laboratory[J]. Proc Natl Acad Sci, 2014, 111(6): 2122-7.

(编辑:吴锦雅)